株価とニュース報道を用いた上場企業の暗黙関係の発見

AUTHOR(S):

馬場, 慧; 馬, 強

CITATION:

馬場, 慧 ...[et al]. 株価とニュース報道を用いた上場企業の暗黙関係の発見. DEIM Forum 2016 論文集 2016: G3-2.

ISSUE DATE:

2016-03

URL:

http://hdl.handle.net/2433/217597

DEIM Forum 2016 G3-2

株価とニュース報道を用いた上場企業の暗黙関係の発見

馬場　慧† 馬　強††

† 京都大学工学部情報学科　〒 606–8501 京都市左京区吉田本町
†† 京都大学大学院情報学研究科　〒 606–8501 京都市左京区吉田本町
E-mail: †baba@db.soc.i.kyoto-u.ac.jp, ††qiang@i.kyoto-u.ac.jp

あらまし　企業間の関係分析は，マーケティングや意思決定において重要である．企業の Web サイトなどで子会社やグループ会社などに関する記述は多いが，スポンサー関係や取引先などの暗黙的に関連する企業に関する情報は少ない．本研究では，関連するニュースイベントに対する株価の動向の類似性を分析して，上場企業間の暗黙的な関係を発見する手法を提案する．提案手法では，まず，株価を市場，業種と企業自身の三つの要因の合成モデルから生成されると仮定し，市場や業種の影響を調整した企業の株価を抽出する．調整済みの株価系列データを正規化した上，関連するニュースイベントの日付を元に実価データの部分系列を抽出し，抽出された部分系列の類似度を計算することで，関連性の強い企業を発見する．東京株式市場の株価データとの財経新聞のニュース記事を用いて提案手法の評価を行う．

キーワード　関係マイニング，投資情報分析，時系列データ

1. はじめに

企業は組織であり，他の組織との関わり合いの中で存続させ成長させることが重要である．企業は他の企業との関係によって多種多様な組織間のネットワークを構成している．このような組織と組織の関係ネットワークを分析する「組織間関係論」は長年研究の対象となっており [1]，現在も研究が行われている．

企業間には様々な関係が存在しており，企業の Web サイト等やニュース記事を見ることによって他の企業との関係を調べることができる．しかし，企業間の関係には様々な種類があり，グループ会社，子会社といった明示的な関係もあれば，スポンサー関係や取引先など直接は明らかにされていない暗黙的な関係もある．企業間の関係性を分析する研究は既に行われているが，インターネットで得られるテキスト情報から明示的な関係のみの分析を行っているものが多い [2]．

本研究では，企業同士の関係を株価という観点から明らかにし，関係のある企業を集めた企業間のネットワークを分析することで，企業の競争力の分析や戦略決定，個人投資家の企業の成長性の分析や投資企業選定の支援を行う．

本研究では意思決定支援に重要である，業績に影響を及ぼす関係を対象とし，株価とニュースイベント情報を併用した手法を提案する．企業間に業績に影響を及ぼす関係が存在すれば，両企業の株価の動きに類似性があると考えられる．企業の業績を調査するに当たって株価は非常に重要な要素の一つであり，株価が上昇しているときは企業の業績もよく，株価が下落しているときは企業の業績も悪いといったことが多い．企業間の株価の動きの関連性を調査することで，その企業間の関係性を明らかにすることもできる．例えば，ある企業が自動車の開発を行った際に別の企業の部品を使用していれば，その部品の売れ行きは自動車の売れ行きに左右される．そのような時，自動車が売れて，開発した企業の業績が良くなり株価が上昇すれば，部品を供給している企業の業績も良くなり株価が上昇する．

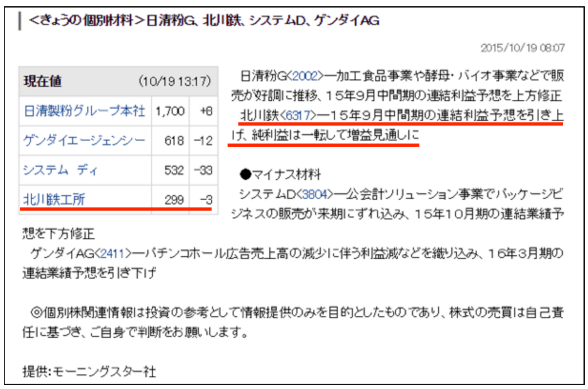

図 1　2015/10/19　MORNINGSTAR より抜粋

企業の株価が上昇するような材料が出ているにも関わらず，株価が下落しているといった場合も存在する．図 1 は 2015 年 10 月 19 日の MORNINGSTAR(注1) で発表されたニュース記事である．北川鉄工所 (6317) は「15 年 9 月中間期の連結利益予想を引き上げ，純利益は一転して増益見通し」という株価が上昇するような材料が出ているが，実際の株価は 3 円のマイナス (前日比-0.993%) になっている．そのような場合に，我々は，企業の業種，市場全体の影響が大きく関係していると考える．以後本論文では純粋な企業自身のみの株価を実価，業種の指数を業種指数，市場の指数を市場指数と呼ぶ．本研究では，企業の株価は実価，業種指数と市場指数からなるとし，株価の合成モデルを提案し，それに基づいて実価を求めて企業の関係を分析する．

企業間の関係を分析する際に，まず，異なる企業の株価の値

(注1)：http://www.morningstar.co.jp/

幅の差を調整するため，企業の実価の前日比を求めて正規化する．次に，企業間の関係が動的に変化することを考慮して，企業に関係するニュースイベントをトリガーとして株価の比較する範囲を決定する．比較範囲の系列データを用いて，企業間の関連度を推定する．本研究では，正規化した実価データを用いて企業間の関連度を計算する手法として，ハミング距離，SAX 法，相関係数と偏相関係数の四つの手法を検討する．

本研究の主な貢献は以下にまとめる．

- 企業の株価の合成モデルを提案し，市場や業種の影響を調整して，より企業自身の業績を反映している実価を用いて企業間の株価の動向の類似性を求める．(3 節)
- 前日比を用いて株価を正規化して類似度を計算する手法を提案している．提案手法は異なる企業間の株価の単価の差を吸収し，値のトレンドにフォーカスした分析ができる．(4 節)
- ニュースイベントを用いて比較する株価の範囲を決める．企業間の関係は時間の経過と共に変化するため，比較対象の株価の範囲の選別が非常に重要である．提案手法は，ニュースイベントをトリガーとし，比較範囲を動的に決めることで，直近のニュースイベントが発生してからの企業間の関係を抽出することが可能である．(4.5 節)

本論文の構成は次の通りである．2 節では企業間の関係や，テキスト情報が株価に与える影響に関する関連研究を示し，3 節では本研究で仮定する株価データの合成モデルについて記す．4 節では企業間の関係を分析する際の手法を説明し，5 節では評価実験の方法とその結果を示す．そして，6 節は本研究のまとめである．

2. 関 連 研 究

企業間ネットワークを抽出，分析し，知見を得る研究は盛んにおこなわれている．金らは Web 上に存在している情報から企業間の関係を明らかにし，企業ネットワークを抽出する手法を提案している [2]．金らの研究は企業間の関係性を導き出す情報として，Web 上のテキスト情報のみを対象としており，抽出する関係性の対象も提携関係と訴訟関係のみに絞っているが，本研究では関係性の対象を取らず，企業の業績に焦点を当てて関係性があるかどうかを判断する．

Woo らは協調性，適応性，雰囲気といった観点から企業間関係の質を評価し，関係とサービスの質の関係を明らかにする手法を提案している [3]．また，Rauyruen らは企業間関係の質を決定する要因としてサービスの品質や売り手へのコミットメント，信頼性，満足度をあげており，関係の質と購買の意図，繰り返し買うかどうかの忠実性に及ぼす影響の関係を調べた [4]．Woo らや Rauyruen らは企業間の関係の質をサービスの向上や顧客分析に利用するものと位置づけており，意思決定には利用しない．

サッカーとファイナンスの関係の分野では，Michael らや Aliakbar らはサッカーのクラブチームの試合結果がスポンサー企業の株価に影響を与えることを明らかにしている [5] [6]．これらの研究では暗黙に関係する企業をスポンサー関係をもとに手動で与えているが，本研究では株価のデータをもとに自動的に発見する．

ニュース記事やソーシャルネットワーク等のテキスト情報が株価に影響を与えるという研究も多く存在する．Tetlock は Wall Street Journal の市場観測のコラム記事から悲観度を抽出し，ダウ工業平均株価と関係していることを明らかにした [7]．また，Bollen らは Twitter のテキスト情報であるツイートを解析し，世間のムードを測ることによってダウ工業平均株価の変動の予測する試みを行っている [8]．このようにテキスト情報は株価の変動を測るうえで重要な指標のひとつとなっている．これらの研究では，テキスト情報を将来的な株価変動の予測に用いているが，本研究では株価変動の予測ではなくテキスト情報を株価の変動のタイミングを示す情報として利用する．

3. 株価データの合成モデル

本節では株価データの合成モデルについて述べる．業界や市場の影響によって変動し，企業そのものの動きを表していない株価のデータで関連度を計算することを避けるため，株価の合成モデルを提案する．さらに，この合成モデルを用いて，企業の株価から市場と業種の影響を調整する手法を提案する．

1 節で説明したように，本研究は，企業自身の業績のみからなる株価を実価，業種の指数を業種指数，市場の指数を市場指数と呼ぶ．実価，業種指数，市場指数が株価を構成している式を求める際に，季節調整 [9] の考え方を用いる．季節調整とは経済統計の時系列データから季節要因を取り除く手法であり，株価の分析にも使用される．本研究では季節調整で利用される合成モデルである乗法モデルを用いる．元の株価のデータを X_t，実価を C_t，業種指数を I_t，市場指数を M_t とすると，株価の合成モデルは下記のように定義される．

$$X_t = C_t \times I_t \times M_t \tag{1}$$

ただし，I_t には調節する企業が属する業種の業種別株価指数，M_t には調節する企業が東証 1 部であれば TOPIX，東証 2 部であれば東証 2 部株価指数といったようにその企業が属する市場の株価指数を用いる．企業が属する業種は，東京証券取引所が定めた 33 業種を利用する．この合成モデルを用いて，調整する企業の株価の実価 C_t を式 (2) のように計算する．

$$C_t = \frac{X_t}{I_t \times M_t} \tag{2}$$

例として，トヨタ自動車 (7203) の 2015 年 12 月 30 日の実価を求める手順を示す．トヨタ自動車は業種として輸送用機器，市場としては東証 1 部に属している．2015 年 12 月 30 日の株価，業種指数，市場指数はそれぞれ 7488，3267.86，1547.3 であるので，トヨタ自動車の実価 (c_{toyota}) は $c_{toyota} = \dfrac{7488}{3267.86 \times 1547.3} = 0.00148$ である．

表 1　データセット

株価データ (株価データダウンロードサイト)	ニュースデータ (財経新聞)	企業情報 (Yahoo! Finance)
個別銘柄の株価	ニュース記事のタイトル	企業コード
業種別株価指数	ニュース記事の本文	企業名
市場ごとの株価指数	ニュース記事の日付	業種名
		市場名

4.　関係分析手法

本節では，株価とニュースを用いた企業間の関係の分析手法について述べる．4.1 節では本研究で扱うデータセットについて説明し，4.2 節では入力データと出力データおよび処理の流れを示す．4.3 節では，問合せの企業とその関連企業候補の株価データの正規化について述べる．4.4 節と 4.5 節では，正規化されたデータを用いて企業間の関連度とそのためのデータ範囲の決める方法についてそれぞれ説明する．

4.1　データセット

本研究で使用する入力は表 1 にまとめる．

株価データは株価データダウンロードサイト (注2) のデータを用いる．株価データダウンロードサイトから 2007 年から現在までの TOPIX，業種別株価指数等の様々な株価指数データ，全ての上場企業の個別銘柄データを csv ファイルでダウンロードして利用する．各株価指数データには株式市場営業日の始値，高値，安値，終値，各個別銘柄データには株式市場営業日の始値，高値，安値，終値，出来高，売買代金が保存されているが，その日にあったニュースイベントの全ての影響を反映している終値を株価のデータとして用いる．各個別銘柄データの中でも，東証 1 部，東証 2 部，東証マザーズ，ジャスダックに上場している企業のうち，Yahoo! Finance(注3) に企業のページが存在する銘柄のみを対象とし，株価データに欠損値を含んでいる銘柄は対象としない．本研究では使用するニュース記事を 2010 年 9 月 14 日以降のものとしているので 2010 年から 2015 年の株価データを使用する．

ニュース記事のデータは財経新聞 (注4) のニュース記事を使用する．財経新聞は様々なカテゴリのニュース記事を配信しているが，本研究では株価に直接影響を与えやすい企業の動きのニュースを多く扱っている企業・産業カテゴリのニュース記事のみを対象とする．対象とするニュース記事の期間は最大にさかのぼることができる 2010 年 9 月 14 日から 2015 年の株式市場最終日である 2015 年 12 月 30 日の 5 年強とする．また，ニュースイベントによって比較範囲を決定した際に，範囲内のデータがすべて揃っていない企業は関連度を計算することが不可能であるため，そのような銘柄は対象としない．

企業名の記事への言及の有無に基づいてその企業の関連ニュースであるか否かを決める．企業名としては Yahoo! Finance で使用されているものを用いる．例えば，企業コード 7203 の企業であれば ‘トヨタ自動車’ であり，企業コード 7267 の企業は ‘ホンダ’ である．また，企業名に全角アルファベット等が含まれている場合には半角に変換したのちに検索する．

(注2)：http://k-db.com/
(注3)：http://finance.yahoo.co.jp/
(注4)：http://www.zaikei.co.jp/

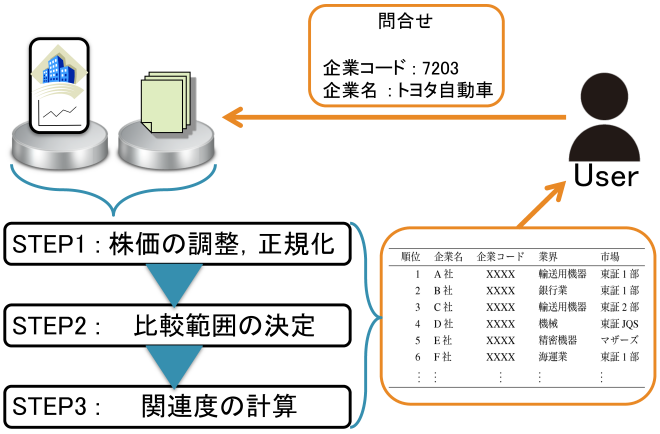

図 2　処理の流れ

4.2　手法の概要

提案手法の処理の流れを図 2 に示す．

- 入力データ

入力データは 4.1 節に説明したように，ニュースデータと株価データを用いる．また，調べたい企業コードをユーザから入力する問合せとする．つまり，入力は問合せである企業コード，企業株価データベースとニュースデータベースからなる．

- 出力データ

出力データとして問合せである上場企業とその他の上場企業の関連度を比較し，高い順にランキング形式で表示する．ユーザには上場企業のランキングを提示し，企業間関係の分析の支援をおこなう．

- 処理

入力データを出力データにする際の処理の流れは以下の通りである．

– STEP1

企業の株価データ，業種指数，市場指数を 3 節の式 (2) を用いて計算し各企業の実価を求め，正規化を行う．

– STEP2

ニュース記事の中から対象としている企業に言及しているニュース記事を取り出し，発表された日付を元に関連度を求める際に用いるデータの範囲を決定する．

– STEP3

実際に，STEP1 で求めた実価と STEP2 で設定した範囲を元に問合せである上場企業とその他すべての上場企業との関連度を計算する．

4.3　実価の正規化

前日比は株価データを解析する際に重要な指標のひとつである．多くのファイナンスの Web サイトで公開していることか

表 2　KDDI の終値

日付	株価 (KDDI)	業種指数 (情報・通信業)	市場指数 (TOPIX)	実価 (KDDI)
2015/3/25	8144	2776.22	1592.01	0.001843
2015/3/26	8186	2751.24	1568.82	0.001897
2015/3/27	2727.0	2729.79	1552.78	0.000643
2015/3/30	2723.0	2737.35	1557.77	0.000639

らもその重要度がうかがえる．前日比は現在の株価が前日の株価の終値の値から何 % 増減しているかを表している．企業によって値の大きさが違う株価をすべて同じスケールのデータに変換できるので，比較することが容易になる．また，ニュースイベントが発生してから株価がどのように推移したのかを見る必要がある本研究では，前日比により前日からの株価の推移を明らかにする．

前日比を求める際，前日のデータが必要であるため，前日のデータが存在しない株取引開始の日の値は 0 とする．式 (3) は前日の株価の終値を cp_{d-1}，当日の株価の終値を cp_d としたときの前日比 (δ_d) を求める式である．

$$\delta_d = \begin{cases} 0, & d = 1 \\ \dfrac{cp_d - cp_{d-1}}{cp_{d-1}} \times 100, & d > 1 \end{cases} \tag{3}$$

株式分割や株式併合が行われた際には δ の絶対値が非常に大きな値となってしまう場合がある．例として，KDDI(9433) の株式分割が行われた時の株価の変動を示す．表 2 は，株式分割が行われた前後 2 日間の株価データである．この時，実価を前日比での正規化を行うとそれぞれの期間の前日比は

$$(-2.844, 194.798, 0.747)$$

となり，株式が分割された日の前日比の値が極端に大きくなっているのがわかる．

そこで，前日や翌日の値に比べて当日の δ の絶対値が極端に大きくなっていることがあれば，外れ値として処理することによって，株式分割や株式併合の際の値の変化に対処する．実価の前日比のデータを $\delta_i (\delta_1 = 0, i = 1, 2, \cdots, n)$ とし，μ を δ_i の平均値，σ を δ_i の標準偏差とすると，δ_i の式 (4)(注5) が成り立つ際に δ_i は外れ値であるとする．外れ値を処理する際の式は式 (5) である．

$$\delta_i \leqq \mu - 10\sigma \ or \ \delta_i \geqq \mu + 10\sigma \tag{4}$$

$$\delta_i = \begin{cases} \delta_{n-1}, & i = n \\ \dfrac{\delta_{i-1} + \delta_{i+1}}{2}, & i \neq n \end{cases} \tag{5}$$

この式により，大きく外れた値を棄却し，その日の前日比としては前後の日の前日比の平均値を与える．

さらに，比較する企業群の中で外れ値を除いた一番大きな δ の絶対値ですべての δ を割ることによって，−1 から 1 の値に正規化する．

4.4　関　連　度

時系列データの関連性分析の手法には様々なものがあり，文字列にエンコードしたのち，文字列の類似度を計算する手法もあれば，そのままのデータを比較していく手法もある．本研究では，時系列データの関連度を求めるにあたって，両方の手法を用いる．文字列にエンコードしたのちに文字列同士の類似度を計算する手法としては 4.4.1 節のハミング距離と 4.4.2 節の SAX 法を使用し，そのままの時系列データの関連度を求める手法としては 4.4.3 節の相関係数と 4.4.4 節の偏相関係数を使用する．実験では，この 4 種類の手法について実験結果を用いて考察を行う．

4.4.1　ハミング距離

同じ長さをもつ文字列同士の類似度を示す尺度としてハミング距離が利用されている．ハミング距離では距離が小さいほど比較する二つの時系列データは類似しているといえ，2 つの文字列の距離は，一方の文字列に文字の置換を行いもう一方の文字列に変形する最小のコストとして求められる．従来のハミング距離の計算では，文字の置換コストはすべて 1 であるが，本研究では文字の辞書順を考慮する．

正規化した実価を文字列にエンコードする際，使用する文字の数 k(注6) はパラメータとして与え，実験的に変更できるようにする．k の値はすべてのデータが同じ文字にエンコードされないように 1 より大きくとる．各文字のとる値の範囲は −1 から 1 までの範囲を k の値で等間隔に割ったものとする．

文字の種類は前日から株価がいかほど上昇下落したかを表すものであるので，文字の置換のコストをすべて 1 にしてしまうと株価の関連度を測る際に不都合が生じる．このような不都合を防ぐために，ハミング距離の文字の置換に文字の差によるコストを付加する．具体的には隣接する文字に置換する場合のコストを 1 とする．a から j への置換のコストは 9 となり，a から b のコストは 1 となるので，文字の種類によってコストが変わる．この新たに文字の置換の定義を変更したハミング距離を用いて文字列同士の距離を求める．2 つの比較する文字列を $A = a_1a_2a_3 \cdots a_n$，$B = b_1b_2b_3 \cdots b_n$ とし，関数 $dict(x)$ を文字 x を辞書順でならべたときに何番目かを表す関数とすると，ハミング距離を求める式は式 (6) のようになる．

$$d_h(A, B) = \sum_{i=1}^{n} |dict(a_i) - dict(b_i)| \tag{6}$$

関連度 ($r_h(a, b)$) を求める式は以下の式である．この値が 0 に近いほど関連度が高いと言える．$len(X)$ は文字列 X の長さ

(注5)：σ の係数である 10 は実験的に決定した値であるので，最適な係数は今後検討していく予定である．

(注6)：最適な k の値を求める方法については，今後の研究で検討する予定である．

を表す関数とする．本研究では，比較する文字列の長さは同じであるので $len(A) = len(B)$ である．以下，関数 $len()$ は同様の定義のものとして扱う．

$$r_h(A,B) = \frac{d_h(A,B)}{len(A)} \tag{7}$$

4.4.2 SAX 法

SAX 法 (Symbolic Aggregate approXimation) は時系列データを分析する際に用いられる手法の一つであり，Lin らによって提案された手法 [10] である．SAX 法も時系列データを文字列にエンコードしてから類似度を測る手法であるが，ハミング距離の際に用いた文字列へのエンコードとは異なり，標準正規分布に従いエンコードを行う．

SAX 法ではまず，PAA(Piecewise Aggregate Approximation) という操作を行う．PAA は時系列データを時間軸に沿って等間隔に w 個のフレームに分割し，各フレームの平均値を求め，各フレームに含まれているデータをそのフレームの平均値に置き換えるという操作である．時系列データ $C = c_1, c_2, \cdots, c_n$ を $\overline{C} = \overline{c}_1, \overline{c}_2, \cdots, \overline{c}_w$ に変換するとき，変換後の値を求める式は式 (8) のようになる．

$$\overline{c}_i = \frac{w}{n} \sum_{j=\frac{n}{w}(i-1)+1}^{\frac{n}{w}i} c_j \tag{8}$$

本研究では 1 日ごとの前日からの実価の上昇下落に注視しているため，平均値を取ると特徴が失われてしまう可能性が高い．そのため，SAX 法では PAA で時系列データを操作してから類似度の計算を行うのに対し，本研究ではフレームの数をニュース発生時から経過した日数とし，1 日ごとの実価のデータを PAA の変換を行った後のデータとして扱う．つまり，実価は 1 日のデータの平均と考える．

PAA で変換した後の実数値のデータはアルファベットの文字列にエンコードされる．データが標準正規分布に従うという仮定のもとエンコードを行うので，文字列の各アルファベットが同一確率で出現する．アルファベットの文字数をアルファベットサイズと呼び，文字の分割点はアルファベットサイズごとに決めておくことができる．例えば，アルファベットサイズが 10 のときの分割点はは $(\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8, \beta_9) = (-1.28, -0.84, -0.52, -0.25, 0, 0.25, 0.52, 0.84, 1.28)$ となる．

$Q = q_1 q_2 \cdots q_w$ と $C = c_1 c_2 \cdots c_w$ を文字列にエンコード後の時系列データとする．Q と C の距離を求めるために，Lee らは式 (9) の距離関数を定義している．

$$d_m(Q,C) = \sqrt{\frac{n}{w}} \sqrt{\sum_{i=1}^{w} (dist(q_i, c_i))^2} \tag{9}$$

ただし，$dict(q,c)$ は以下の通りの定義により計算される関数である．

$$cell_{r,c} = \begin{cases} 0, & |r-c| \leqq 1 \\ \beta_{max(r,c)-1} - \beta_{min(r,c)}, & otherwise \end{cases} \tag{10}$$

ただし，$cell_{r,c}$ は辞書順に並べたときの r 番目と c 番目のアルファベット間の距離を表す．アルファベットサイズが 10 のときを例にとって考えると，a と e の距離は $dict(a,e) = cell_{1,5} = \beta_4 - \beta_1 = -0.25 - (-1.28) = 1.03$ となる．

関連度 $(r_s(Q,C))$ を求める式は以下の式である．この値が 0 に近いほど関連度が高いとする．

$$r_s(Q,C) = \frac{d_m(Q,C)}{len(Q)} \tag{11}$$

4.4.3 相関係数

相関係数は 2 つの変数データの相関の程度を示す数値であり，統計学の相関分析の分野で広く用いられている．相関係数を求める際に様々な手法があるが，本研究ではピアソンの積率相関係数を用いる．

2 つのデータ列を $X = x_1, x_2, \cdots, x_n$ と $Y = y_1, y_2, \cdots, y_n$ とすると，ピアソンの積率相関係数 r は X と Y の共分散と X と Y それぞれの標準偏差で求めることができる．相関係数 $cor(X,Y)$ を求める式は式 (12) のようになる．

$$cor(X,Y) = \frac{\sum_{i=1}^{n}(x_i - \overline{X})(y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{Y})^2}} \tag{12}$$

ただし，$\overline{X}$，$\overline{Y}$ はそれぞれ X，Y の相加平均を表している．

関連度 $(r_c(X,Y))$ を求める式は以下の式である．この値が 1 に近いほど関連度が高いと言える．$lendata(X)$ はデータ X の個数を表している．以下 $lendata()$ は同様の定義として扱う．本研究では，扱うデータの個数は同じなので，$lendata(X) = lendata(Y)$ である．

$$r_c(X,Y) = \frac{cor(X,Y)}{lendata(X)} \tag{13}$$

4.4.4 偏相関係数

相関係数では 2 つの変数の相関を明らかにするものであるが，2 つの変数に影響を与える他の変数が存在する可能性を考慮していない．このような影響を与える変数が存在していて，本来相関がないような 2 つの変数が相関があるような結果が出てしまうことを疑似相関といい，相関関係を調査する際に考慮すべきものとなっている．偏相関係数は 2 つの変数の間に存在する相関関係を求める際に，その他の影響を与えている変数の影響を取り除いた相関の程度を示す数値である．

本研究で扱うデータである実価は時系列データなので，両方が時間に関係している変数である．よって，2 つの実価から時間という変数の影響を除外した相関を求める．2 つの実価を X，

Y，時間を T とおき，$cor(X,Y)$，$cor(X,T)$，$cor(Y,T)$ をそれぞれ変数間の相関係数とすると，偏相関係数 $pcor(X,Y,T)$ は式 (14) で求められる．

$$pcor(X,Y,T) = \frac{cor(X,Y) - (cor(X,T) \times cor(Y,T))}{\sqrt{1 - cor(X,T)^2}\sqrt{1 - cor(Y,T)^2}} \tag{14}$$

関連度 ($r_p(X,Y,T)$) を求める式は以下の式である．この値が 1 に近いほど関連度が高いと言える．このとき，$lendata(X) = lendata(Y) = lendata(T)$ である．

$$r_p(X,Y,T) = \frac{pcor(X,Y)}{lendata(X)} \tag{15}$$

4.5 ニュースイベントによる比較範囲の決定

企業同士の関連性は時間の経過とともに遷移する．企業間の関係を調べるときに，対応する時間の範囲を決める必要がある．時間の範囲を決定することにより，企業間の関係をダイナミックに捉え，データセットに用意されている全ての期間を比較する必要がなくなるので計算量も小さくすることができる．

例えば，ある企業が他の企業に部品供給をしている場合を考える．納品先の企業が部品供給する企業をコンペティション方式で決めているとすると，コンペティションに勝利して部品供給をしている間は企業同士の業績はリンクしており，関連度は高くなると考えられるが，別の企業がコンペティションに勝利してしまうと，その瞬間から部品供給が打ち切られるので関連度が低くなってしまうと予想される．

企業の関係を調査するにあたって過去の企業同士の関係も重要であるが，意思決定を行う際には今現在の関係性を考慮すべきである．2 節で述べたように，ニュースイベントは株価の上昇下落を判断するうえで必要不可欠な材料であり，発生して株価が大きく変動する可能性が高い．よって 2015 年 12 月 30 日現在，一番最近発生したニュースイベントから現在までの範囲を比較範囲とし，実価の比較を行うことで，最新の企業間同士の関連度を抽出することができる．

ニュース記事は収集できる最も過去のニュース記事が発表された 2010 年 9 月 14 日以降のものとし，$t = \tau$ のときにニュースイベントが発生したとする．期間 $[\tau, 2015/12/30]$ の間隔が十分に大きければその時間帯に対応する実価データを比較することで関連度を計算し，関連性の推移を測る．しかし，本研究で用いる時系列データの比較手法はデータ数が少なすぎれば関連度が大きいか全くなしかの二極化してしまうため，時系列データにある程度の長さがなければ関連度を導出することができない．そこで，期間 $[\tau, 2015/12/30]$ の間隔に使用するデータの長さの最小値を設定しなければならない (注7)．本研究では，使用するデータの長さの最小値として 30 日間と設定することにする．つまり，式 (16) がいつでも成り立つものとする．

(注7)：最適な最小値を求める方法について今後検討する予定である．

表 3　評価実験における問合せ企業の一覧

企業名	企業コード	業種	市場
JT	2914	食料品	東証 1 部
セブン&アイ・ホールディングス	3382	小売業	東証 1 部
武田薬品工業	4502	医薬品	東証 1 部
トヨタ自動車	7203	輸送用機器	東証 1 部
ホンダ	7267	輸送用機器	東証 1 部
キヤノン	7751	電気機器	東証 1 部
三菱 UFJ フィナンシャル・グループ	8306	銀行業	東証 1 部
みずほフィナンシャルグループ	8411	銀行業	東証 1 部
日本電信電話	9432	情報・通信	東証 1 部
NTT ドコモ	9437	情報・通信	東証 1 部

$$(2015/12/30 - \tau) + 1 \geqq 30 \tag{16}$$

この期間に起こったニュースイベントは現在の関連度に関するニュースイベントとする．決定された比較範囲を用いて比較手法を適応する．

5. 評価実験

5.1 実験の概要

株価の合成モデルおよびそれを用いた企業間の関係分析手法を評価するための実験を行った．実験では，被験者が実際に調べてつけた企業の関連度のランキングと提案手法によって導出された企業の関連度のランキングを比べて nDCG(Normalized Discounted Cumulative Gain) [11] を計算する．被験者として 4 人の大学 (院) 生に対し，10 社の問合せ企業と提案手法でランクインしたそれぞれの問合せ企業との関連の強い企業群を与え，企業間の関係の強さを 5 段階評価してもらい，その結果に基づいて提案手法と比較手法の nDCG の値をそれぞれ算出した．関連度を求める際に使用する 4 つの手法で得られたランキングの nDCG を計算し比較を行う．次に，合成モデルを用いて株価調整を行った場合のランキングと調整を行わない場合のランキング結果の nDCG を比較し，合成モデルの有用性について考察する．

5.2 データセット

評価実験において，問合せとなる対象企業を 2016 年 1 月 18 日当時時価ランキングを元に上位 10 社とした．また，業種による違いも考慮するため，一つの業種に対して最大 2 社までとし，ニュースイベントが発生してから 2015 年 12 月 30 日のデータが揃っていない企業は対象外とする．対象とする 10 社は表 3 にまとめる．ニュース記事は 2010 年 9 月 14 日から 2015 年 12 月 30 日までの財経新聞の記事を用いた．

5.3 パラメータの設定

企業の実価データを文字列にエンコードする際に用いられる文字数 k の値 (4.4.1 節を参照) を決定する調査を行った．調査では計算量の観点からも考えて，k の値を $k = 10$ から $k = 10^2$

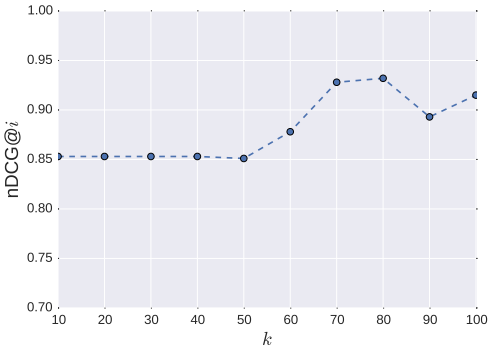

図 3　k の nDCG@30

までの間に存在する k の値を 10 刻みで変動させ，2016 年 1 月 18 日当時時価総額 1 位のトヨタ自動車 (7203) を問合せの対象企業として関連企業のランキング結果に基づいて k を選定した．

k の値一つに対して 4.4.1 節で述べた手法を用いてランキングを上位 30 件まで導出し，各 k の値で導出された企業の OR をとって企業の和集合を作成する．ランキングの評価を行う際には nDCG を用いる．その企業の和集合に被験者一人が企業間の業績の関係を 5 段階で評価した後，k の値ごとに nDCG@30 を計算し，nDCG の値が最大となったものをハミング距離での k の値として使用する．結果のグラフを図 3 に示す．

図 3 から，$k = 80$ のときに nDCG 最大となる．k の値が小さい時はランキングに出てくる企業が変化していないことから，k の値はある程度大きい方がよいと思われる．今回の実験では，ハミング距離を用いる際に nDCG の値が最大だった $k = 80$ とするが，最適な k の値を決める方法については今後検討していく予定である．

5.4　評 価 方 法

評価方法としては，まず，各関連度計算手法を用いて対象とする企業とその他の企業とのランキングの上位 15 件を求める．次に，各手法で求められた企業群の OR をとり，4 手法のいずれかにランクインした企業の和集合を作成する．和集合に含まれる企業と問合せとなる対象企業の業績の関係性が大きいと思われるものから順に 5 点，4 点，3 点，2 点，1 点というように 5 段階で 4 人の被験者に評価してもらった．人間が与えた関係の強さのランキングと手法で計算されたランキングの各手法の nDCG を算出して比較することで，関連度の計算手法の精度を比較する．

nDCG の値は問合せとする対象企業 10 社の値の平均とする．また，その値が最も大きかった手法を使用し，合成モデルを用いて株価調整を行った場合のランキングと調整を行わない場合のランキングの nDCG を比較する．nDCG は上位 5 件，10 件と 15 件での値を考慮する．

5.5　実 験 結 果

5.5.1　合成モデルとそれに基づく株価調整手法の評価

実価の nDCG@$i(i = 5, 10, 15)$ の各値をプロットし，グラフ化したものが図 4 である．調整前の株価の nDCG@$i(i = 5, 10, 15)$ の各値をプロットし，グラフ化したものが図 5 である．横軸 i が評価するランキングの長さ，縦軸が nDCG@i の値を表している．図 4 から文字列にエンコードし，文字列同士の距離を計算する手法 (ハミング距離，SAX 法) より，文字列にエンコードすることなくデータ同士の相関係数を計算する手法 (相関係数，偏相関係数) が適していることがわかる．

表 4　nDCG の平均値

	ハミング距離	SAX 法	相関係数	偏相関係数
実価	0.865	0.857	0.902	0.903
株価	0.844	0.837	0.844	0.838

図 4 と図 5 を比較して，nDCG@i の i の値が小さい時には nDCG の大きな差は見られないが，i の値が大きくなると図 5 では nDCG の値が急落しているのが見て取れる．このことから，nDCG@i の i の値を大きくすれば，調整を行っていない株価データを比較するよりも，提案したモデルで比較を行った方が企業の関連度をより正確に捉えられることが分かる．

図 4 と図 5 の nDCG@$i(i = 5, 10, 15)$ の値の平均値をとった値を表 4 にまとめる．表 4 より，相関係数よりも偏相関係数の手法がより適していると言える．また，全ての手法において，調整を行った実価の nDCG の平均値が調整していない株価の nDCG の平均値よりも大きくなっており，提案した株価モデルが効果的であるとわかる．

5.5.2　ニュースイベントを用いた比較範囲の選定手法の評価

ニュースイベントによって比較する範囲を決定した場合の偏相関係数の nDCG@$i(i = 5, 10, 15)$ と比較する範囲を限定せずに 2010 年 9 月 14 日以降の両企業の株価のデータが存在している範囲全てで比較した場合の偏相関係数の nDCG@$i(i = 5, 10, 15)$ を表したグラフが図 6 である．なお，実験で用いた企業の中には比較範囲を広げると欠損値を含んでいる企業も存在したため，株価データに 2 日連続で欠損値を含んでいる企業については比較を行わないものとし，それ以外の株価データに欠損値を含んでいる企業については，元の株価データを $x_i(i = 1, 2, \cdots, n)$ とし，欠損値 (x_j) を以下の式 (17) に基づいて補完した．

$$x_j = \begin{cases} x_2, & j = 1 \\ x_{n-1}, & j = n \\ \dfrac{x_{j-1} + x_{j+1}}{2}, & otherwise \end{cases} \tag{17}$$

図 6 から，ニュースイベントを使い，範囲を限定して比較した方が全期間を比較範囲としたときより nDCG@i の値が大きいことがわかる．このことから，ニュースイベントを活用することで，計算量が減少するだけでなく，企業間の関係性として正しいランキング順に並べることができると考えられる．

5.6　議　　論

この節では問合せの対象企業ごとの結果を調査する．表 5 は nDCG@15 のときの結果である．

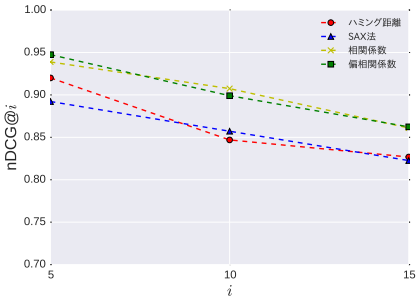

図 4　実価の nDCG@i

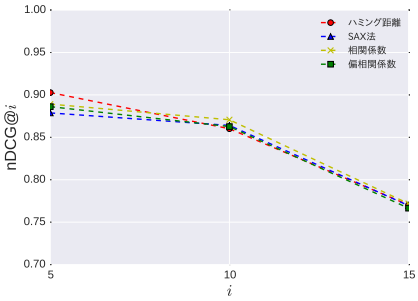

図 5　株価の nDCG@i

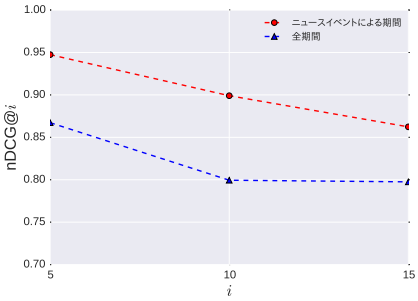

図 6　比較範囲ごとの nDCG@i の値

表 5　nDCG@15

企業コード	ハミング距離	SAX 法	相関係数	偏相関係数
2914	0.872	0.847	0.830	0.834
3382	0.695	0.718	0.850	0.856
4502	0.868	0.798	0.806	0.804
7203	0.878	0.916	0.917	0.920
7267	0.847	0.856	0.894	0.894
7751	0.811	0.729	0.838	0.832
8306	0.804	0.819	0.866	0.867
8411	0.912	0.936	0.915	0.916
9432	0.772	0.798	0.776	0.806
9437	0.809	0.809	0.919	0.895

表 5 において相関係数と偏相関係数の値はかなり似通った値となっている．これは実価という時系列データは時間に比例して増えるなどといったデータではないため，時間の変数の影響はほとんどないものと考えられる．

nDCG@i($i = 5, 10, 15$) の平均をとった結果としては偏相関係数の値が一番高いが，企業によっては文字列にエンコードして文字列同士の距離を計算する手法が有用である企業もある．武田薬品工業 (4502) はハミング距離の nDCG@15 が最大となっている．i の値にもよるが，ハミング距離や SAX 法の nDCGi が高くなっている企業も存在する．よって，エンコード手法を改良して他の距離関数に基づく関連度の計算手法について今後検討したい．

特に，編集距離では置換の他にも削除と挿入の操作を定義しているので，時系列データのずれに対応することができる．本研究では株価のデータとして終値を使用しているので 1 日のずれというのは大きいかもしれないが，株価データの粒度を大きくした時などはずれを考慮することによって，企業の業績推移から関連企業への業績の推移が始まるまでのタイムラグの影響を排除することができると考えられる．

6. おわりに

本研究では，株価の合成モデルとそれに基づく，株価とニュース報道を併用した企業の関係分析手法の提案を行っている．提案手法は企業間の関連度の強さをランキング形式で明らかにし，企業や個人投資家などの企業間ネットワーク分析の支援を行う効果が期待できる．

評価実験において，上場企業 10 社の関連企業を提案手法で算出する関連度でランキングし，nDCG の結果を用いて評価を行った．実験結果から，株価の合成モデルによる株価の調整とニュースイベントによる比較範囲の選定の有効性を確認した．また，ケーススタディの結果では，株価を文字列にエンコードして編集距離で関連度を計算することの可能性が示された．

今後，クラウドソーシングを用いた大規模な評価実験の実施，パラメータ k の設定，比較範囲の最小値の決定，暗黙な関係発見の評価や応用システムの構築について行う予定である．

謝　辞

本研究の一部は，科研費 (課題番号 25700033) と SCAT 研究費助成による．

文　献

[1] 山倉健嗣: 組織間関係と組織間関係論, 横浜経営研究, Vol. 16, No. 2, pp. 166–178 (1995).
[2] 金英子, 松尾豊, 石塚満: Web 上の情報を用いた企業間関係の抽出, 人工知能学会論文誌, Vol. 22, pp. 48–57 (2007).
[3] Woo, K., Ennew, C. T.: Business‐to‐business relationship quality: An IMP interaction‐based conceptualization and measurement, *European Journal of Marketing*, Vol. 38, pp. 1252–1271 (2004).
[4] Rauyruen, P. and Miller, K. E.: Relationship quality as a predictor of B2B customer loyalty, *Journal of business research*, Vol. 60, No. 1, pp. 21–31 (2007).
[5] Hanke, M. and Kirchler, M.: Football championships and jersey sponsors' stock prices: an empirical investigation, *The European Journal of Finance*, Vol. 19, pp. 228–241 (2012).
[6] Ramezani, A., Mardani, H., Emamgholipour, M. and Mardani, S.: The Effect of the Results of Football Champions League Games on Sponsors' Stock Prices: Evidence from Iran, *World Applied Sciences Journal*, Vol. 20, pp. 102–106 (2012).
[7] Tetlock, P. C.: Giving Content to Investor Sentiment: The Role of Media in the Stock Market, *The Journal of Finance*, Vol. 62, pp. 1139–1168 (2007).
[8] Bollen, J., Mao, H. and Zeng, X.: Twitter mood predicts the stock market, *Journal of Computational Science*, Vol. 2, pp. 1–8 (2011).
[9] 有田帝馬: 入門 季節調整, 東洋経済新報社 (2012).
[10] Lin, J., Keogh, E., Lonardi, S. and Chiu, B.: A Symbolic Representation of Time Series, with Implications for Streaming Algorithms, *DMKD '03*, pp. 2–11 (2003).
[11] Järvelin, K. and Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques, *ACM TOIS*, Vol. 20, No. 4, pp. 422–446 (2002).